2007

Sampling within households in household surveys

Robert Graham Clark
University of Wollongong, rclark@uow.edu.au

David G. Steel
University of Wollongong, dsteel@uow.edu.au

Sampling within households in household surveys

Abstract

The number of people to select within selected households has significant consequences for the conduct and output of household surveys. The operational and data quality implications of this choice are carefully considered in many surveys, but the impact on statistical efficiency is not well understood. The usual approach is to select all people in each selected household, where operational and data quality concerns make this feasible. If not, one person is usually selected from each selected household. We find that this strategy is not always justified, and develop intermediate designs between these two extremes. Current practices were developed when household survey field procedures needed to be simple and robust, however more complex designs are now feasible due to the increasing use of computer-assisted interviewing. We develop more flexible designs by optimising survey cost, based on a simple cost model, subject to a required variance for an estimator of population total. The innovation lies in the fact that household sample sizes are small integers, which creates challenges in both design and estimation. The new methods are evaluated empirically using census and health survey data, showing considerable improvement over existing methods in some cases.

Keywords

Cost–variance optimization, Health surveys, Household surveys, Model-assisted

Disciplines

Physical Sciences and Mathematics

Publication Details

This paper was originally published as: Clark, RG, Sampling within households in household surveys, Journal of the Royal Statistical Society Series A, 2007, 170 (1), 63-82. Published by Blackwell on behalf of the Royal Statistical Society 2007. The definitive version is available at www.blackwell-synergy.com.

Sampling Within Households in Household Surveys

Robert G. Clark and David G. Steel [1]

Abstract

The number of people to select within selected households has significant consequences for the conduct and output of household surveys. The operational and data quality implications of this choice are carefully considered in many surveys, but the impact on statistical efficiency is not well understood. The usual approach is to select all people in each selected household, where operational and data quality concerns make this feasible. If not, one person is usually selected from each selected household. We find that this strategy is not always justified, and develop intermediate designs between these two extremes. Current practices were developed when household survey field procedures needed to be simple and robust, however more complex designs are now feasible due to the increasing use of computer-assisted interviewing. We develop more flexible designs by optimising survey cost, based on a simple cost model, subject to a required variance for an estimator of population total. The innovation lies in the fact that household sample sizes are small integers, which creates challenges in both design and estimation. The new methods are evaluated empirically using census and health survey data, showing considerable improvement over existing methods in some cases.

Key Words: sample design, model-assisted, survey estimation, household surveys, cost-variance optimisation

1. Introduction

Household surveys are one of the most important tools used by national statistical agencies, social scientists and market researchers to measure characteristics of human populations. The basic methodology is to select a sample of households, possibly geographically clustered, then to collect data on some or all people in the household. Data may also be collected on the household itself. The most

[1]Centre for Statistical and Survey Methodology, University of Wollongong, NSW 2522 Australia. E-mail: Robert_Clark@uow.edu.au. This work was jointly supported by the Australian Research Council and the Australian Bureau of Statistics.

common modes of collection are face-to-face and telephone interviewing, possibly using computer assisted interviewing (CAI). Internet surveys are also becoming popular.

There is surprisingly little literature on how many people should be selected from selected households. Factors affecting this choice include:

(1) *The impact on response rates.* Collecting data on more people in the household would be perceived as more burdensome for the respondents. This is particularly the case when the interview is lengthy or sensitive.

(2) *The amount of survey content that can be collected.* If multiple people are selected from selected households, the respondent burden can be reduced by shortening the interviews or removing sensitive questions. This reduces the amount of information that can be produced from the survey.

(3) *Data quality.* Collecting data from multiple people in each household may mean that sensitive questions are not answered accurately. Alternatively, the answers of the first respondent may influence the answers of subsequent respondents in the same household.

(4) *Collection method.* Sometimes one household member can report on behalf of others in the household. This is called proxy interviewing. Anyone for whom data is collected is deemed to be in the sample regardless of whether the data was collected directly or by proxy. Proxy interviewing can make

sampling all people in the household cheaper, although the total interview time is usually not much reduced.

(5) *Selection method.* When one person is selected from each selected households, this selection can inadvertently be biased towards people who are more cooperative or more often at home. Techniques such as Kish Grids (e.g. Kish, 1967, pp.398-401) involve listing all household members before selecting one at random.

(6) *Complexity of analysing survey data.* The sample design has implications for analysis of survey data. If all household members are selected, then statistical models may have to incorporate dependencies between values in the same household. Alternatively, if one/household sampling is used, the probability of selection will depend on the household size, and this may have to be reflected in the analysis method.

(7) *Intrinsic interest in Dependencies within Households.* There may be substantive interest in dependencies between values for different people in the same household. (For example, do unemployed people tend to live together? How does parents' health influence their children's health?) In this case, multiple, preferably all, people should be selected in selected households.

(8) *The relative cost of sampling an additional household compared with sampling an additional person.* If the costs associated with selecting an addi-

tional household (e.g. travel costs, cost of repeated attempts to contact a household) are high, then a clustered design with many selections in each household (e.g. selecting all people in the household) is appropriate. The converse is also true. See, for example, Foreman (1991, p.216).

(9) *The intracluster correlation (R).* This is a measure of how similar values are for different people in the same household. If R is high, then additional interviews in the same household do not add much information, so that a less clustered design (e.g. one person per household) is appropriate.

In practice, all/household sampling is usually adopted unless (1), (2) or (3) prevent this. Foreman (1991, p.396) argued that "for each sample dwelling, all of the households and all of the persons eligible under the survey coverage rules would be included in the survey ... this is usually advisable on grounds of sampling efficiency and cost ... This arrangement may not be desirable, however, when respondents are subject to lengthy and perhaps uncomfortable interviews to which entire households might object, and when the reaction of a household member to interviewing might prejudice the response of others. In such cases a subsample of one household member is selected at random".

We find, however, that selecting one person per household, or an intermediate option, can be more statistically efficient than selecting all people in a household, even when the latter is operationally feasible.

We use the framework of cost-variance optimisation (e.g. Hansen et al., 1953,

chapters 6 and 7; Foreman, 1991, chapter 8; Lohr, 1999, section 5.5), where the cost according to a simple cost model is minimised subject to a variance constraint (or the variance is minimised for fixed cost), to evaluate and compare alternative within-household sample sizes. Applying this framework to household surveys requires new methods, because the number of people within each household is a small integer, typically between 1 and 4 if the survey scope is restricted to adults.

It will be assumed that: a sample of m households is selected using simple random sampling without replacement (SRSWOR); people within selected households are also selected using SRSWOR; the only information available to help choose the within-household sample sizes is the household sizes for the selected households; and the regression estimator (e.g. Sarndal et al., 1992) is used to estimate the population total of a person-level variable of interest.

The assumption of SRSWOR of households is a simplification. More complex designs are often used; for example there may be an initial, possibly stratified, stage of selection of geographic areas or banks of telephone numbers before the selection of households and people. The extension to these more complex first stage sample designs would be straightforward by introducing a design effect for the household stage of selection into the variance expressions in Sections 3 and 4 (Kish, 1967). We have omitted this step to simplify presentation.

The assumption of SRSWOR of people within households reflects the common practice in household surveys and many other two-stage surveys.

The assumption that only the household size can be used to guide the within-

household sample size reflects the current practice in household surveys, where either "all per household" or "one per household" designs predominate. In one per household designs, the probability of selection of a person is inversely proportional to the household size, leading to variability in estimation weights, which in turn inflate the variance of estimators (see Silva & Skinner, 1997 for the effect of variation in estimation weights in general and Clark & Steel, 2000 for one/household sampling in particular). So it makes sense to make use of the household size to alleviate this problem, at least partly. Furthermore, information on the household size is usually obtained as part of the normal course of survey operations.

It would be possible to collect more detailed information on the household and use this to guide the selection of individuals. This would complicate survey operations and create a more complex sample design problem. The simple case we have assumed gives useful gains in efficiency and forms a sensible starting point for more complex survey designs which could be developed in the future.

This article uses the model-assisted framework for inference, where expectation and variance are over repeated sampling from a fixed population, and the regression estimator is used to estimate population totals (Sarndal et al., 1992). These expectations and variances are sometimes called the design expectation and the design variance. Section 2 defines the main notation for the paper.

An obvious solution to the sample design problem we have posed is to minimise the cost for fixed variance with respect to integer values of within-household

sample sizes. It is assumed that household sizes of selected households are the only information available prior to administering the full survey. This suggests making the within-household sample sizes a function of the household size. Section 3 develops this integer allocation approach.

A more sophisticated solution is to allow the within-household sample sizes to be different for each household. This would enable the average within-household sample size for households of a given size to be a non-integer, enabling better control over the level of clustering of the sample. Section 4 develops this "fractional allocation" approach, including the complex issue of how to construct estimators for this design. Section 5 is an empirical study and Section 6 contains conclusions.

2. Notation

The population of clusters, indexed by g, is denoted U_1, of size M. The sample of m clusters is s_1. The population of units, indexed by i, is denoted U, of size N. The sample of n units is s. For household surveys as defined here, clusters are households and units are people.

The probability of selection for unit i is $\pi_i = P[i \in s]$. The population of N_g units in cluster g is U_g and the sample of n_g units selected from cluster g is s_g. The mean cluster size is $\bar{N} = \frac{N}{M} = \frac{1}{M}\sum_{g\in U_1} N_g$ and the maximum cluster size is $A = max\{N_g : g \in U_1\}$. The mean within-cluster sample size is $\bar{n} = \frac{n}{m} = \frac{1}{m}\sum_{g\in s_1} n_g$.

It will be assumed that m and M are large and that m is much smaller than M, but that N_g may be small. All approximations will be based on dropping

terms of order m^{-1}, M^{-1} or m/M relative to the remainder of the expression.

The variable of interest for unit i is y_i. The aim is to estimate $Y = \sum_{i \in U} y_i$. Typically there is some auxiliary information about the whole population which can be used to enhance estimation of Y. Let $\boldsymbol{x}_i$ be the set of auxiliary variables for unit i, and let $\boldsymbol{X} = \sum_{i \in U} \boldsymbol{x}_i$. The regression estimator of Y is

$$\hat{Y}_r = \sum_{i \in s} d_i \left(y_i - \boldsymbol{b}^T \boldsymbol{x}_i \right) + \boldsymbol{b}^T \boldsymbol{X} \tag{1}$$

where

$$\boldsymbol{b} = \left(\sum_{i \in s} d_i c_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \sum_{i \in s} d_i c_i \boldsymbol{x}_i y_i$$

is a weighted least squares regression coefficient of y_i on $\boldsymbol{x}_i$, and where d_i are "initial weights" and c_i are weights usually chosen based on a variance model for y_i. The usual approach is to set $d_i = \pi_i^{-1}$, however any set of weights can be used provided that they are unbiased weights in the sense that $(\sum_{i \in s} d_i y_i)$ is unbiased for Y for any variable of interest.

If the initial weights are unbiased, then $\hat{Y}_r$ is approximately equal to

$$\tilde{Y}_r = Y + \sum_{i \in s} d_i e_i$$

where $e_i = y_i - \boldsymbol{B}^T \boldsymbol{x}_i$ and $\boldsymbol{B} = \left(\sum_{i \in U} c_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \sum_{i \in U} c_i \boldsymbol{x}_i y_i$ is a population weighted least squares regression coefficient of $\{y_i\}$ on $\{\boldsymbol{x_i}\}$. (Sarndal et al., 1992 show this for $d_i = \pi_i^{-1}$ in chapter 6 and for an alternative d_i for the case of two-phase sampling in chapter 9). This approximation can be used to derive approximate variances for $\hat{Y}_r$.

We will assume that the weights c_i used in the calculation of regression parameters have the property that $c_i = \boldsymbol{\lambda}^{\boldsymbol{T}}\boldsymbol{x_i}$ for all $i \in U$, for some vector $\boldsymbol{\lambda}$. In this case, the population mean, $\bar{E}$, of $\{e_i\}$, is zero (this can be shown using the same argument as in Sarndal et al., 1992, Result 6.5.1, p. 231). This condition simplifies a number of our results, and would usually be satisfied in practice. For example it is true if the ratio estimator is used, or if $c_i = 1$ and the auxiliary variables include an element equal to 1 for all i.

We write $e_{g1} = \sum_{i \in U_g} e_i$ for the cluster totals of e_i and $\bar{e}_g = N_g^{-1} e_{g1}$ for the cluster means. The variance for cluster g is $S_g^2 = (N_g - 1)^{-1} \sum_{i \in U_g} (e_i - \bar{e}_g)^2$.

The population mean is $\bar{E} = \frac{1}{N} \sum_{i \in U} e_i = 0$. The population mean of the cluster totals is $\bar{E}_1 = \frac{1}{M} \sum_{g \in U_1} e_{g1} = \bar{N}\bar{E} = 0$. The population variance is $S^2 = \frac{1}{N-1} \sum_{i \in U} \left(e_i - \bar{E}\right)^2$. The population variance of the cluster totals is $S_1^2 = \frac{1}{M-1} \sum_{i \in U} \left(e_{g1} - \bar{E}_1\right)^2$. The intra-cluster correlation is $R = 1 - \frac{S_W^2}{S^2}$ where $S_W^2 = \frac{1}{N-M} \sum_{g \in U_1} \sum_{i \in U_g} (e_i - \bar{e}_g)^2$. R measures how similar the values of e_i are for units in the same cluster.

We will make use of population parameters for clusters of size a, for $a = 1, \ldots, A$. The population of clusters of size a is U_{1a} (of size M_a) and the population of units in these clusters is U_a (of size $N_a = aM_a$). The sample size of clusters of size a is m_a and a total of n_a units within these clusters are selected.

For clusters of size a, the mean within-cluster sample size is $\bar{n}_a = n_a/m_a$, the mean of e_i is $\bar{E}_a = N_a^{-1} \sum_{i \in U_a} e_i$, the mean of e_{g1} is $\bar{E}_{1a} = M_a^{-1} \sum_{g \in U_{1a}} e_{g1}$, and

the intra-class correlation is $R_a = 1 - \frac{S^2_{Wa}}{S^2_a}$ where $S^2_a = \frac{1}{N_a - 1} \sum_{i \in U_a} \left(e_i - \bar{E}_a \right)^2$ and $S^2_{Wa} = \frac{1}{N_a - M_a} \sum_{g \in U_{1a}} \sum_{i \in U_g} (e_i - \bar{e}_g)^2$. A useful identity which follows directly from these definitions is that the mean of S^2_g over $g \in U_{1a}$ is $(1 - R_a) S^2_a$.

The between-cluster variance is

$$S^2_B = \bar{N}^{-2} M^{-1} \sum_{g \in U_1} \left(e_{g1} - \bar{E}_1 \right)^2 = \bar{N}^{-2} M^{-1} \sum_{g \in U_1} e^2_{g1}.$$

The between-cluster variance for clusters of size a is

$$S^2_{Ba} = a^{-2} M_a^{-1} \sum_{g \in U_{1a}} \left(e_{g1} - \bar{E}_{1a} \right)^2 = M_a^{-1} \sum_{g \in U_{1a}} \left(\bar{e}_g - \bar{E}_a \right)^2.$$

A standard ANOVA decomposition gives $S^2_a \approx (N_a - M_a) S^2_{Ba} + M_a S^2_{Wa}$ so $S^2_{Ba} = S^2_a a^{-1} (1 + (a-1) R_a)$. We can decompose S^2_B as

$$\begin{aligned} S^2_B &= \bar{N}^{-2} M^{-1} \sum_{g \in U_1} e^2_{g1} = \bar{N}^{-2} M^{-1} \sum_{g \in U_1} N^2_g \bar{e}^2_g \\ &= \bar{N}^{-2} M^{-1} \sum_{a=1}^{A} a^2 \sum_{g \in U_{1a}} \left\{ \bar{e}_g - \bar{E}_a + \bar{E}_a \right\}^2 \\ &= \bar{N}^{-2} M^{-1} \sum_{a=1}^{A} a^2 \left\{ M_a S^2_{Ba} + M_a \bar{E}^2_a \right\} \\ &= \sum_{a=1}^{A} \left(\frac{a}{\bar{N}} \right)^2 \frac{M_a}{M} \left(S^2_a a^{-1} (1 + (a-1) R_a) + \bar{E}^2_a \right). \end{aligned} \tag{2}$$

3. Fixed Integer Allocations

In this section it is assumed that the within-cluster sample size is a function of the cluster size: $n_g = c_a$ if $N_g = a$, where c_a are integers between 1 and a. The aim is to find the best values of m and c_a for $a = 1, \ldots, A$.

3.1 Variance

The design variance of $\tilde{Y}_r$ is approximately equal to

$$V \approx \frac{M^2}{m}\left(1-\frac{m}{M}\right)\bar{N}^2 S_B^2 + \frac{M}{m}\sum_{g\in U_1}\frac{N_g^2}{n_g}\left(1-\frac{n_g}{N_g}\right)S_g^2 \tag{3}$$

(Hansen et al., 1953, equation 5.3, p.317 with only one stratum). For $g \in U_{1a}$, $n_g = c_a$ and $N_g = a$. Hence

$$\begin{aligned}
V &\approx \frac{M^2}{m}\left(1-\frac{m}{M}\right)\bar{N}^2 S_B^2 + \frac{M}{m}\sum_{a=1}^{A}\sum_{g\in U_{1a}}\frac{a^2}{c_a}\left(1-\frac{c_a}{a}\right)S_g^2 \\
&\approx \frac{M^2}{m}\bar{N}^2 S_B^2 + \frac{M}{m}\sum_{a=1}^{A}\frac{a^2}{c_a}\left(1-\frac{c_a}{a}\right)M_a S_a^2\left(1-R_a\right) \\
&\approx \frac{M^2}{m}\bar{N}^2\sum_{a=1}^{A}\frac{M_a}{M}\left(\frac{a}{\bar{N}}\right)^2\left\{S_a^2 a^{-1}\left(1+(a-1)R_a\right)+\bar{E}_a^2\right\} + \frac{M^2}{m}\sum_{a=1}^{A}\frac{M_a}{M}\left(\frac{a^2}{c_a}-a\right)S_a^2\left(1-R_a\right) \\
&\approx \frac{M^2}{m}\bar{N}^2\sum_{a=1}^{A}\frac{M_a}{M}\left(\frac{a}{\bar{N}}\right)^2\left(S_a^2 R_a+\bar{E}_a^2\right) + \frac{M^2}{m}\sum_{a=1}^{A}\frac{M_a}{M}\frac{a^2}{c_a}S_a^2\left(1-R_a\right)
\end{aligned}$$

It is convenient to write V in terms of m and $n_{0a} = E\left[n_a\right]$ which for this design is equal to $n_{0a} = \frac{M_a}{M}mc_a$. Let

$$V_1 = M^2\sum_{a=1}^{A}\frac{M_a}{M}a^2\left(S_a^2 R_a+\bar{E}_a^2\right) \tag{4}$$

$$V_{2a} = M_a^2 a^2 S_a^2\left(1-R_a\right). \tag{5}$$

Then

$$\begin{aligned}
V &\approx m^{-1}V_1 + m^{-1}M\sum_{a=1}^{A}c_a^{-1}M_a^{-1}V_{2a} \\
&\approx m^{-1}V_1 + \sum_{a=1}^{A}n_{0a}^{-1}V_{2a}.
\end{aligned}$$

3.2 Cost

The cost of implementing the sample design is assumed to be $C = mC_1 + \sum_{a=1}^{A} C_{2a} n_a$. This cost expression is not suitable for optimising the sample design, because it depends on the values of n_a, which are not known in advance of sampling. The expected cost must be used instead:

$$C_E = E[C] = mC_1 + \sum_{a=1}^{A} C_{2a} n_{0a} \tag{6}$$

which for the fixed integer design becomes

$$C_E = mC_1 + \sum_{a=1}^{A} C_{2a} m \frac{M_a}{M} c_a. \tag{7}$$

Linear cost models of this form are commonly used by official statistics agencies. A linear cost model is often adequate for the purpose of sample design, even though it cannot perfectly capture the real cost structure. For example, when one adult reports on behalf of others in the household, the cost per interview may be higher for the first interview than for additional interviews. More complex cost models have sometimes been used (e.g. Csenki, 1997).

3.3 Optimal Allocation Ignoring Integer Effects

Minimising C_E subject to fixed variance $V = V_f$ with respect to m and $\{c_a\}$ is equivalent to minimising with respect to m and $\{n_{0a}\}$. The optimal values are

$$\begin{aligned} m &= V_f^{-1} \sqrt{V_1/C_1} \left(\sqrt{V_1 C_1} + \sum_{a=1}^{A} \sqrt{V_{2a} C_{2a}} \right) \\ n_{0a} &= V_f^{-1} \sqrt{V_{2a}/C_{2a}} \left(\sqrt{V_1 C_1} + \sum_{a=1}^{A} \sqrt{V_{2a} C_{2a}} \right) \end{aligned}$$

(e.g. Csenki, 1997) and hence

$$c_a = \frac{n_{0a}}{mM_a/M} = \frac{M}{M_a}\sqrt{\frac{V_{2a}/C_{2a}}{V_1/C_1}} \tag{8}$$

The optimal cluster sample sizes c_a do not depend on the variance constraint V_f.

These expressions for the optimal c_a and m will generally have non-integer values. In practice they could be rounded to the nearest integers. If c_a and m are large then this rounding may have little effect, but c_a would typically be small for household surveys so that rounding would have a significant impact. One effect of rounding is that the cost constraint is not met precisely. In this case, m could be re-calculated to meet the cost constraint given the set of integers c_a obtained by rounding (8).

3.4 Optimal Integer Allocation

A better procedure is to find the set of integers $\{c_a\}$ and the value m which minimise C_E subject to $V = V_f$. Given a set of values of c_a, m is determined by the variance constraint:

$$\begin{aligned} V_f &= V_1 m^{-1} + \sum_{a=1}^{A} V_{2a} n_{0a}^{-1} = V_1 m^{-1} + \sum_{a=1}^{A} V_{2a} \left(m \frac{M_a}{M} c_a \right)^{-1} \\ \rightarrow m &= V_f^{-1} \left\{ V_1 + \sum_{a=1}^{A} V_{2a} \frac{M}{M_a} c_a^{-1} \right\}. \end{aligned}$$

For this value of m, the expected cost is

$$\begin{aligned} C_E &= C_1 m + \sum_{a=1}^{A} C_{2a} n_{0a} = m \left(C_1 + \sum_{a=1}^{A} C_{2a} \frac{M_a}{M} c_a \right) \\ &= V_f^{-1} \left\{ V_1 + \sum_{a=1}^{A} V_{2a} \frac{M}{M_a} c_a^{-1} \right\} \left\{ C_1 + \sum_{a=1}^{A} C_{2a} \frac{M_a}{M} c_a \right\} \end{aligned} \tag{9}$$

Hence the optimal c_a can be found by minimising (9) with respect to $\{c_a : a = 1, \ldots, A\}$ over $c_a \in \{1, \ldots, a\}$. If A is not too large this can be done by calculating (9) for every possible set of $\{c_a : a = 1, \ldots, A\}$ since there are only $A!$ possibilities. Expression (9) shows that the optimal values of c_a do not depend on the variance constraint, which is convenient because it means that alternative within-cluster designs can be compared without specifying V_f.

4. Fractional Allocations

4.1 The Basic Design

In Section 3, the mean within-cluster sample sizes $\bar{n}_a$ were equal to the integers c_a as the n_g were equal to c_a for each $g \in s_{1a}$. In this section, the $\bar{n}_a$ are allowed to be non-integer, by allowing n_g to have different integer values for each $g \in s_{1a}$. Non-integer $\bar{n}_a$ may give lower cost for fixed variance by allowing greater control over the probabilities of selection and the level of clustering of the sample.

When a selected household is first contacted, the size of the household is collected. No other auxiliary information is available to distinguish between selected households, prior to administering the full interview. Therefore it is reasonable to randomly assign values of n_g for $g \in s_{1a}$ from an integer-valued distribution depending on a. Any distribution of integers can be used, however the design is much simpler to understand and calculate if n_g takes on the values of two neighbouring integers, or a single integer. In this case, the distribution of n_g is fully specified by its expected value. This simplification is justifiable on practical grounds, and also turns out to be the optimal strategy (Clark, 2002). Once n_g is

generated, the sample s_g is selected by SRSWOR conditional on n_g.

The value of n_g can be generated as part of the field process. In a face-to-face interview, this would be done on the doorstep, after collecting information on the number of household members, but before conducting the interview. This would be straightforward in a survey using computer-assisted interviewing (CAI), because household data would be entered into the computer as it was reported. Because the value of n_g would be generated live in the field, the values of n_g would necessarily be independent for different g because coordination between interviewers would not be feasible.

The expected value of n_g will be denoted θ_a for $N_g = a$. Let $[\theta_a]$ be the non-integer part of θ_a, let θ_a^- be highest integer less than or equal to θ_a, and let $\theta_a^+ = \theta_a^- + 1$. So n_g is equal to θ_a^+ with probability $[\theta_a]$ and to θ_a^- with probability $(1 - [\theta_a])$. We use $\boldsymbol{n}$ to denote the vector of n_g over $g \in s_1$.

4.2 Estimation for Fractional Allocations

The fact that the n_g are randomly generated leads to several choices of estimator, because there is a choice whether realized or expected values of n_g are used in weighting. We will discuss three options, recommend a weighting method, and state the variance of the recommended estimator.

All three options are special cases of the regression estimator (1) with different choices of the initial weights d_i. The first option is

$$d_i \;=\; \pi_i^{-1} = \left\{ \frac{m}{M} P\left[i \in s \,|s_1 \right] \right\}^{-1} = \left\{ \frac{m}{M} E\left[P\left[i \in s \,|s_1, \boldsymbol{n} \right] |s_1 \right] \right\}^{-1}$$

$$
\begin{aligned}
&= \left\{ \frac{m}{M} E\left[\frac{n_g}{N_g} \,|s_1 \right] \right\}^{-1} \\
&= \frac{M}{m} \frac{a}{\theta_a} \qquad (10)
\end{aligned}
$$

for $i \in U_g$ and $N_g = a$. These weights depend on θ_a, the expected value of n_g, but not on the realized values of n_g.

The second option is

$$d_i = \frac{M}{m} \frac{N_g}{n_g} \qquad (11)$$

for $i \in U_g$. These weights are unbiased because

$$
\begin{aligned}
E\left[\sum_{i \in s} d_i y_i \right] &= E\left[E\left[\sum_{i \in s} y_i | s_1, \boldsymbol{n} \right] \right] \\
&= E\left[\sum_{g \in s_1} \frac{M}{m} E\left[\frac{N_g}{n_g} \sum_{i \in s_g} y_i | s_1, \boldsymbol{n} \right] \right] \\
&= E\left[\sum_{g \in s_1} \frac{M}{m} y_{g1} \right] = Y.
\end{aligned}
$$

The third option is to replace θ_a in (10) by the realized average within-household sample size $\bar{n}_a$:

$$d_i = \frac{M}{m} \frac{a}{\bar{n}_a}. \qquad (12)$$

These weights are also unbiased.

Proof: $\bar{n}_a$ is the mean of the n_g over $g \in s_{1a}$, and the n_g are independent and identically distributed dichotomous variables within s_{1a}. Hence $E\left[n_g \,|\bar{n}_a\right] = \bar{n}_a$ and $P\left[i \in s \,|s_1, \bar{n}_a\right] = \frac{\bar{n}_a}{a}$ so that $E\left[\sum_{i \in s_g} \frac{N_g}{n_g} y_i \,|s_1, \bar{n}_a\right] = y_{g1}$.

It is not obvious which of (10), (11) or (12) should be used. Method (11) is attractive because it gives a sensible weight when each household is looked at

individually: $\frac{N_g}{n_g}\sum_{i\in s_g} y_i$ is a sensible estimator of y_{g1}. This would be expected to be beneficial if households vary substantially, which occurs when across-household variability of y_i is high, which occurs when R is close to 1. The weights in (10) and (12) are less intuitive for estimating y_{g1} individually, but these weights vary less across the sample, which may lead to lower variances. Clark (2002) proved the following results about the three methods:

(i) The regression estimator based on (12) has asymptotic variance less than or equal to that based on (10) (as m and M but not N_g tend to infinity). The improvement depends on the extent to which the $\bar{E}_a$ vary over $a = 1, \ldots, A$.

(ii) The optimal choice of d_i (where d_i may depend on n_g but not any values of the variable of interest) is a nonlinear interpolation between (11) and (12). The interpolation depends on R and θ_a, with (11) being optimal when $R = 1$, and (12) being optimal when $R = 0$.

We do not propose using the optimal d_i in general, because of their complexity and the inconvenient property that weights are different for different variables of interest, depending on their value of R. Instead, (12) with $d_i = \frac{M}{m}\frac{a}{\bar{n}_a}$ is recommended, since R is closer to 0 than 1 for most variables of interest. This estimator will be denoted by $\hat{Y}_r^*$. The initial weights d_i are unbiased so $\hat{Y}_r^* \approx \tilde{Y}_r^*$ where $\tilde{Y}_r^* = Y + \sum_{i\in s} d_i e_i$. The theorem below states the variance of $\tilde{Y}_r^*$ which is the approximate variance of $\hat{Y}_r^*$.

Theorem 1: Suppose that clusters are selected by SRSWOR and that units

within cluster are selected by SRSWOR conditional on n_g, where the n_g are independently generated by an integer-valued distribution on $\{0, ..., a\}$ for $N_g = a$. Let $E[n_g] = \theta_a > 0$ and $var[n_g] = \gamma_a^2$, for $N_g = a$. Let $n_{0a} = E[n_a] = m\frac{M_a}{M}\theta_a$. Then

$$var\left[\tilde{Y}_r^*\right] \approx m^{-1}V_1 + \sum_{a=1}^{A} n_{0a}^{-1}V_{2a}^* \tag{13}$$

where V_1 is defined as in Section 3.1, and

$$V_{2a}^* = M_a^2 a^2 S_a^2 \left(1 - R_a + \gamma_a^2 \theta_a^{-1} R_a\right).$$

See Appendix for Sketch of Proof. Full details are contained in Clark (2002).

Theorem 1 applies even if n_g can be zero, provided that $\theta_a > 0$. Notice that V_{2a}^* is itself a function of θ_a.

If the n_g are not random, then $\gamma_a^2 = 0$ and $V_{2a}^* = V_{2a}$, so that the variance in Theorem 1 is equal to the variance for the integer allocation in Section 3.1. If the n_g are random and only take on the values of θ_a^- and θ_a^+, for $N_g = a$, then

$$\gamma_a^2 = [\theta_a](1 - [\theta_a]) > 0 \tag{14}$$

so there is a penalty to the variance of $\hat{Y}_r^*$ because $V_{2a}^* > V_{2a}$ in this case. However, the penalty can be worthwhile, because it enables designs which are less clustered than the all per household design and have less variable selection probabilities than the one per household design. This will be shown empirically in Section 5.

4.3 Optimal Fractional Allocation

The aim is to minimise the expected cost, C_E, for fixed variance $V = V_f$ where V is given by Theorem 1. Given a set of values of θ_a (which also determine γ_a^2

and V_{2a}^*), m is determined by the variance constraint:

$$\begin{aligned} V_f &= V_1 m^{-1} + \sum_{a=1}^{A} V_{2a}^* n_{0a}^{-1} = V_1 m^{-1} + \sum_{a=1}^{A} V_{2a}^* \left(m \frac{M_a}{M} c_a \right)^{-1} \\ \rightarrow m &= V_f^{-1} \left\{ V_1 + \sum_{a=1}^{A} V_{2a}^* \frac{M}{M_a} c_a^{-1} \right\} \end{aligned} \tag{15}$$

For this value of m, the expected cost is

$$\begin{aligned} C_E &= C_1 m + \sum_{a=1}^{A} C_{2a} n_{0a} \\ &= C_1 m + \sum_{a=1}^{A} C_{2a} \left(m \frac{M_a}{M} \theta_a \right) \\ &= V_f^{-1} \left\{ C_1 + \sum_{a=1}^{A} C_{2a} \frac{M_a}{M} \theta_a \right\} \left\{ V_1 + \sum_{a=1}^{A} V_{2a}^* \frac{M}{M_a} \theta_a^{-1} \right\} \end{aligned} \tag{16}$$

Hence the optimal θ_a can be found by minimising (16) over $0 < \theta_a \leq a$.

Expression (16) gives the variance as a function of θ_a. This function can be minimized over $0 < \theta_a \leq a$ to give optimal values of $\{\theta_1, \ldots, \theta_A\}$. From the form of (14) it appears that V_{2a}^* and hence (16) are discontinuous at integer values of θ_a. In fact (16) is continuous but not differentiable at these points; details are omitted here. These non-differentiable points explain why the optimal θ_a are sometimes exactly equal to integers in the numerical study. Once the optimal θ_a have been calculated, m is determined by (15).

We attempted to use the R routine NLMINB to perform the numerical optimisation. However, the package had difficulties due to the non-differentiability of (16) at integer values of θ_a. We obtained good results by optimising over $\theta_a \in [c_a, c_a + 1]$ for every set of integers $\{c_a\}$ such that $c_a \in \{0, ..., a\}$. There are $A!$ such sets, but this was quite feasible for $A = 6$. The optimum was then

obtained by using the best of the $A! = 720$ optimisations.

5. Empirical Study

A range of household sample designs was evaluated empirically using data from the 1% sample unit record file of the 1991 Australian Census of Population and Housing and the 1995 Australian National Health Survey. Households of size 6 or higher were excluded, representing less than 0.2% of the population. Only people aged 18 and over were included. After these exclusions, the census data included 74938 households and the health data included 20548 households. The values of $\bar{E}_a$, S_a^2, R_a, S_a^2 etc were estimated from sample data for 4 census variables and 3 health variables. Table 1 contains basic descriptive information on these variables. The values of M_a/M were estimated from the census dataset and were equal to 33.2%, 48.8%, 12.0%, 4.7%, 1.0% and 0.2% for $a = 1, ..., 6$.

Table 2 shows the optimal integer design and the optimal fractional design for each variable. This table assumes $C_{1a} = C_1$ and $C_{2a} = C_2$ with $C_1/C_2 = 0.1$. That is, the costs associated with selecting an additional household are one-tenth that of selecting an additional person. This could be the case for a telephone survey with a particularly efficient redialling methodology. The values of the within-household sample size, c_a, are shown for the optimal integer design for each variable and $a = 1, \ldots, 6$. These are followed in brackets by the expected within-household sample size (θ_a) for the optimal fractional design.

The last column of Table 2 contains $E[\bar{n}]$ which is the expected value of the average within household sample size across all selected households. The first

value in this column is for the optimal integer design, and the bracketed value following it is for the optimal fractional design. $E[\bar{n}]$ is a simple overall measure of the level of clustering of the design.

Tables 3 and 4 are similar to Table 2, but show different cost regimes. Table 3 has $C_1/C_2 = 0.25$ which could arise in many telephone surveys. Table 4 has $C_1/C_2 = 0.5$ which would be more typical of face-to-face surveys.

In Tables 2, 3 and 4, c_a and θ_a are both increasing with a. This makes sense because it means that the probabilities of selection ($\frac{m}{M}\frac{c_a}{a}$ and $\frac{m}{M}\frac{\theta_a}{a}$) are fairly constant across the population. This is particularly noticeable for the fractional design where θ_a/a is almost constant over $a = 1, ..., 4$ (which covers over 90% of all households). The fractional design can achieve more equal selection probabilities than the integer design, because c_a is bounded below by 1 (so that $c_a/a = 1$ for $a = 1$) whereas θ_a can be less than 1.

It would generally be expected that variables with low values of R will have a more clustered optimal design (i.e. high values of $\bar{n}$). The converse also applies. This follows from (8) noting that R is closely related to V_1/V_{2a}. This is sometimes but not always true in Tables 2, 3 and 4. Arthritis, which has the second lowest value of R, always has the most clustered optimal designs. Unemployment, which has an even lower value of R, was very clustered for $C_1/C_2 = 0.5$ in Table 4, but was not particularly clustered for C_1/C_2 equal to 0.1 and 0.25 in Tables 2 and 3. This is because R is not the only factor mediating the effects of different levels of clustering. When cluster sizes vary significantly, as they do when clusters are

households, the variation in cluster size and the covariation of cluster size with the variable of interest, are also important factors. This is particularly an issue for Unemployment although not for Arthritis (Clark & Steel, 2000: see particularly the discrepancy between "δ_2" and "δ_4" in Table 1).

Comparing Tables 2, 3 and 4 shows that the within-household sample size is increasing with C_1/C_2, for both the integer and fractional designs. This makes sense: if households are more expensive to contact, then the sample should be more clustered and concentrated into less households.

Tables 5, 6 and 7 show the efficiency of the various designs, for C_1/C_2 equal to 0.1, 0.25 and 0.5, relative to the all per household design. The cost saving is shown for fixed variance. The half/household design is the integer design where c_a is equal to $a/2$ rounded up.

In Table 5, the all/household design is more efficient than one/household for unemployment and arthritis (the least clustered variables). The one/household design is more efficient for the other 5 variables. In Table 6, with $C_1/C_2 = 0.25$, all/household is more efficient for 3 of the 7 variables. In Table 7, with $C_1/C_2 = 0.5$, all/household is more efficient for all variables.

The optimal integer design gives cost savings (relative to all/household) of 0.6% to 21.6% for $C_1/C_2 = 0.1$, with a median cost saving of 9.9% across the 7 variables. The median cost saving from the optimal integer design was 6.4% for $C_1/C_2 = 0.25$ but only 1.6% for $C_1/C_2 = 0.5$. The optimal integer design was particularly efficient relative to all/household for the variable Language Difficul-

ties. This variable had the highest intraclass correlation of $R = 0.35$.

In practice, values of θ_a less than 1 would be problematical because this implies subsampling households after they are contacted and the household size ascertained. Tables 5, 6 and 7 also show the optimal fractional allocation subject to an additional constraint that θ_a is greater than or equal to 1. This design performed only slightly better than the optimal integer design. This shows that most of the benefit of the optimal fractional design comes from the ability to subsample contacted households, particularly those of size 1. This can be thought of as a special case of two-phase sampling.

Most surveys collect information on many variables and are designed to meet multiple objectives. The optimal integer and fractional designs described in Tables 2 through 7 are optimal for a given variable, but would usually be suboptimal for other variables to some extent. The half/household design evaluated in Tables 4 through 6 is an attempt to find a compromise design which would perform well across most variables. Most of the optimal integer designs are fairly close to having c_a proportional to a, subject to the need for c_a to be integer, so the half/household design is a sensible approach. Tables 4 through 6 show that this design gives good cost savings relative to all/household for almost all variables, except for arthritis (where R is quite low at 0.15) and for unemployment (but only for $C_1/C_2 = 0.5$). Apart from these cases, the half/household design did very nearly as well as the optimal integer design.

6. Conclusions

This article investigated whether there are better designs than all/household sampling, even when it is feasible to collect data for all household members. We found that all/household can be improved upon if the cost of enumerating each household is less than half of the cost of enumerating each person. This would be the case for many telephone surveys. If $C_1/C_2 = 0.1$, the one/household design is more cost-efficient than all/household, with a median saving of 5.2% across the 7 variables we considered. If $C_1/C_2 = 0.25$ or $C_1/C_2 = 0.1$, the new designs developed in this paper give significant improvements over both all/household and one/household sampling.

The first new design is the optimal integer design, where the within-household sample size is an integer-valued function of the household size. This design gave median cost savings of 9.9% for $C_1/C_2 = 0.1$ and 6.4% for $C_1/C_2 = 0.25$. Given the very large budget of many household surveys, these savings would often be more than enough to justify some additional complication in the survey design.

One difficulty with the optimal integer design is that it is variable-specific, which creates challenges for multipurpose designs. The half/household design, where half of all householders (rounding up) are selected, had virtually the same median cost saving as the optimal integer design. This design is therefore a good general purpose design which is worth considering in any situation in which C_1/C_2 is 0.5 or less.

It is also possible to allow the within-household sample sizes to be a mixture

of integers for households of a given size. This "fractional allocation" approach allows the mean within-household sample size to be non-integer for households of a given size. We found empirically that this design can give quite substantial improvements for $C_1/C_2 = 0.1$, but this was mainly due to the subsampling of single person households after the initial contact. This feature could create confusion amongst interviewers and respondents, and would not be acceptable in the great majority of surveys. This finding suggests research into designs where a subset of contacted single person households are given a reduced questionnaire.

There are many qualitative issues to consider in choosing a within-household sampling approach, including response rate, data quality, complexity of analysis, amount of information which can be collected, cost and variance. These issues need to be considered on a case by case basis for every new survey. The approach we have adopted gives a range of designs and a method of evaluating the cost and variance performance of each design. The qualitative issues can then be included to give an evaluation of the total survey quality for each alternative design. Choosing from this wide range of options and considering cost, variance and qualitative issues, household surveys can be designed to achieve the best quality possible with the resources available.

References

Clark, R. G. (2002). *Sample Design and Estimation for Household Surveys* [PhD Thesis]. University of Wollongong.

Clark, R. G., & Steel, D. G. (2000). The effect of using household as a sampling unit. *International Statistical Review*, *70*(2), 289–314.

Csenki, A. (1997). Optimum allocation in stratified random sampling via Hölder's inequality. *Statistician*, *46*, 439–441.
Foreman, E. (1991). *Survey sampling principles*. New York: Marcel Dekker.
Hansen, M., Hurwitz, W., & Madow, W. (1953). *Sample Survey Methods and Theory Vol.1*. New York: Wiley.
Kish, L. (1967). *Survey sampling*. New York: Wiley.
Lohr, S. L. (1999). *Sampling: Design and analysis*. Duxbury Press.
Sarndal, C., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
Silva, P. L. N., & Skinner, C. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, *23*, 23–32.

Appendix: Sketch of Proof of Theorem 1

$$\begin{aligned} var\left[\tilde{T}_r^*\right] &= var\left[\sum_{i\in s} d_i e_i\right] = varE\left[\sum_{i\in s} d_i e_i|s_1\right] + Evar\left[\sum_{i\in s} d_i e_i|s_1\right] \\ &= \frac{M^2}{m}\bar{N}^2 S_B^2 + Evar\left[\sum_{i\in s} d_i e_i|s_1\right]. \end{aligned} \tag{17}$$

We can express $\sum_{i\in s} d_i e_i$ as

$$\sum_{i\in s} d_i e_i = \frac{M}{m}\sum_{a=1}^{A}\frac{a}{\bar{n}_a}\sum_{g\in s_{1a}}\sum_{i\in s_g} e_i = \sum_{a=1}^{A}\hat{N}_{a\pi 1}\frac{\hat{E}_\pi}{\hat{N}_{a\pi}}$$

where

$$\begin{aligned} \hat{N}_{a\pi 1} &= \frac{M}{m}m_a a = \text{inverse probability estimator of } N_a \text{ using } s_1; \\ \hat{N}_{a\pi} &= \frac{M}{m}\frac{a}{\theta_a}\sum_{g\in s_1} n_g = \text{inverse probability estimator of } N_a \text{ using } s; \\ \hat{E}_{a\pi 1} &= \frac{M}{m}\frac{a}{\theta_a}\sum_{g\in s_1}\sum_{i\in s_g} e_i = \text{inverse probability estimator of } E_a \text{ using } s. \end{aligned}$$

A Taylor series expansion around $\hat{N}_{a\pi 1} = N_a, \hat{N}_{a\pi} = N_a, \hat{E}_{a\pi} = E_a$ gives:

$$\sum_{i\in s} d_i e_i \approx \sum_{a=1}^{A}\sum_{g\in s_{1a}}\frac{M}{m}\frac{a}{\theta_a}\sum_{i\in s_g}\left(e_i - \bar{E}_a\right) + (\text{ terms depending on } s_1 \text{ but not } s \text{ or } \boldsymbol{n}).$$

Noting that $E\left[n_g|s_1\right]=\theta_a$ and $var\left[n_g|s_1\right]=\gamma_a^2$ we find:

$$\begin{aligned}
& var\left[\sum_{i\in s} d_i e_i|s_1\right] = E\left[var\left[\sum_{i\in s} d_i e_i|s_1,\boldsymbol{n}\right]|s_1\right] + var\left[E\left[\sum_{i\in s} d_i e_i|s_1,\boldsymbol{n}\right]|s_1\right] \\
\approx\ & E\left[var\left[\sum_{a=1}^{A}\sum_{g\in s_{1a}}\frac{M}{m}\frac{a}{\theta_a}\sum_{i\in s_g}\left(e_i-\bar{E}_a\right)|s_1,\boldsymbol{n}\right]|s_1\right] \\
& +var\left[E\left[\sum_{a=1}^{A}\sum_{g\in s_{1a}}\frac{M}{m}\frac{a}{\theta_a}\sum_{i\in s_g}\left(e_i-\bar{E}_a\right)|s_1,\boldsymbol{n}\right]|s_1\right] \\
=\ & E\left[\sum_{a=1}^{A}\frac{M^2}{m^2}\frac{a^2}{\theta_a^2}\sum_{g\in s_{1a}} n_g\left(1-\frac{n_g}{N_g}\right)S_g^2|s_1\right] + var\left[\sum_{a=1}^{A}\sum_{g\in s_{1a}}\frac{M}{m}\frac{a}{\theta_a}n_g\left(\bar{e}_g-\bar{E}_a\right)|s_1\right] \\
=\ & \sum_{a=1}^{A}\frac{M^2}{m^2}\frac{a^2}{\theta_a^2}\sum_{g\in s_{1a}}\left\{\theta_a-a^{-1}\left(\theta_a+\gamma_a^2\right)\right\}S_g^2 + \sum_{a=1}^{A}\sum_{g\in s_{1a}}\frac{M^2}{m^2}\frac{a^2}{\theta_a^2}\gamma_a^2\left(\bar{e}_g-\bar{E}_a\right)^2 \\
=\ & \sum_{a=1}^{A}\frac{M^2}{m^2}\frac{a^2}{\theta_a^2}\left\{\theta_a-a^{-1}\left(\theta_a+\gamma_a^2\right)\right\}\sum_{g\in s_{1a}}S_g^2 + \sum_{a=1}^{A}\frac{M^2}{m^2}\frac{a^2}{\theta_a^2}\gamma_a^2\sum_{g\in s_{1a}}\left(\bar{e}_g-\bar{E}_a\right)^2
\end{aligned} \tag{18}$$

Taking the expectation of (18) gives

$$\begin{aligned}
Evar\left[\sum_{i\in s} d_i e_i|s_1\right] &= \sum_{a=1}^{A}\frac{M}{m}\frac{a^2}{\theta_a^2}\left\{\theta_a-a^{-1}\left(\theta_a+\gamma_a^2\right)\right\}\sum_{g\in U_{1a}}S_g^2 \\
&\quad +\sum_{a=1}^{A}\frac{M}{m}\frac{a^2}{\theta_a^2}\gamma_a^2\sum_{g\in U_{1a}}\left(\bar{e}_g-\bar{E}_a\right)^2 \\
&= \sum_{a=1}^{A}\frac{M}{m}\frac{a^2}{\theta_a^2}\left\{\theta_a-a^{-1}\left(\theta_a+\gamma_a^2\right)\right\}M_a S_{Wa}^2 + \sum_{a=1}^{A}\frac{M}{m}\frac{a^2}{\theta_a^2}\gamma_a^2 M_a S_{Ba}^2
\end{aligned} \tag{19}$$

The identities for S_{Wa}^2 and S_{Ba}^2 in terms of S_a^2 and R_a from Section 2 can then be substituted into (19). Then $Evar\left[\sum_{i\in s} d_i e_i|s_1\right]$ can be substituted into (17). Expression (2) for S_B^2 from Section 2 can also be substituted into (17). The result follows from straightforward algebraic manipulation.

Table 1: Descriptive Information on Variables in Study

Variable	Description	$\bar{Y}$	$S/\bar{Y}(\%)$	R
employment	1 for employed persons, 0 for others	0.56	75.8	0.27
unemployment	1 for unemployed persons, 0 for others	0.043	465	0.14
income	annual income in Australian dollars	24200	83.6	0.26
language difficulties	1 if speaks English not well or not at all, otherwise 0	0.030	565	0.35
arthritis	1 if suffers from any type of arthritis, 0 otherwise	0.19	183	0.15
health fair or poor	1 if self-reported health is fair or poor, 0 otherwise	0.069	363	0.27
smoker	1 if currently a smoker, 0 otherwise	0.24	176	0.30

Table 2: $c_a(\theta_a)$ for Optimal Integer (Fractional) Design for $C_1/C_2 = 0.1$

Variable	a=1	a=2	a=3	a=4	a=5	a=6	$E[\bar{n}]$
employment	1 (0.37)	1 (0.76)	2 (1.00)	2 (1.42)	3 (2.09)	3 (3.44)	1.19 (0.71)
unemployment	1 (0.51)	1 (0.92)	2 (1.57)	3 (2.00)	3 (2.93)	4 (3.00)	1.24 (0.94)
income	1 (0.42)	1 (0.89)	2 (1.00)	2 (1.50)	3 (2.00)	3 (2.29)	1.19 (0.79)
language difficulties	1 (0.28)	1 (0.60)	2 (1.00)	3 (1.79)	4 (2.01)	4 (4.86)	1.25 (0.63)
arthritis	1 (0.62)	2 (1.00)	3 (1.50)	3 (1.78)	4 (2.30)	5 (3.04)	1.86 (0.99)
health fair or poor	1 (0.66)	1 (1.00)	2 (1.59)	2 (1.94)	3 (3.00)	3 (2.70)	1.19 (1.03)
smoker	1 (0.53)	1 (1.00)	2 (1.38)	2 (2.00)	3 (2.53)	3 (3.44)	1.19 (0.96)

Table 3: $c_a(\theta_a)$ for Optimal Integer (Fractional) Design for $C_1/C_2 = 0.25$

Variable	a=1	a=2	a=3	a=4	a=5	a=6	$E[\bar{n}]$
employment	1 (1.00)	1 (1.00)	2 (2.00)	2 (2.00)	3 (3.00)	4 (4.00)	1.19 (1.19)
unemployment	1 (0.63)	1 (1.00)	2 (2.00)	3 (2.72)	4 (3.60)	4 (4.20)	1.25 (1.11)
income	1 (0.54)	1 (1.00)	2 (1.39)	2 (2.00)	3 (2.48)	3 (2.92)	1.19 (0.96)
language difficulties	1 (0.35)	1 (0.74)	2 (1.00)	3 (2.00)	4 (2.47)	5 (3.08)	1.25 (0.73)
arthritis	1 (1.00)	2 (1.90)	3 (2.68)	3 (3.08)	4 (4.14)	5 (4.67)	1.86 (1.78)
health fair or poor	1 (0.74)	1 (1.00)	2 (1.82)	2 (2.00)	3 (3.00)	3 (3.10)	1.19 (1.08)
smoker	1 (1.00)	1 (1.00)	2 (2.00)	2 (2.00)	3 (3.00)	4 (4.00)	1.19 (1.19)

Table 4: $c_a(\theta_a)$ for Optimal Integer (Fractional) Design for $C_1/C_2 = 0.5$

Variable	a=1	a=2	a=3	a=4	a=5	a=6	$E[\bar{n}]$
employment	1 (1.00)	1 (1.00)	2 (2.00)	3 (3.00)	3 (3.00)	4 (4.00)	1.24 (1.24)
unemployment	1 (0.95)	2 (1.64)	3 (3.00)	4 (4.00)	5 (5.00)	6 (6.00)	1.92 (1.73)
income	1 (0.65)	1 (1.00)	2 (1.87)	2 (2.00)	3 (2.87)	3 (3.14)	1.19 (1.06)
language difficulties	1 (1.00)	1 (1.00)	2 (2.00)	3 (3.00)	4 (4.00)	5 (5.00)	1.25 (1.25)
arthritis	1 (1.00)	2 (2.00)	3 (3.00)	4 (3.47)	5 (4.65)	5 (5.27)	1.92 (1.89)
health fair or poor	1 (0.85)	1 (1.00)	2 (2.00)	2 (2.03)	4 (3.88)	3 (3.09)	1.20 (1.15)
smoker	1 (1.00)	1 (1.00)	2 (2.00)	3 (3.00)	3 (3.00)	4 (4.00)	1.24 (1.24)

Table 5: Cost Saving (%) of Different Designs for $C_1/C_2 = 0.1$

Variable	One/HH	Opt. Integer	Half/HH	Opt. Fractional	Fractional ($n_g \geq 1$)
employment	6.7	12.9	12.9	19.4	12.9
unemployment	-8.4	6.3	5.9	10.4	6.3
income	5.7	9.5	9.5	16.4	10.0
language difficulties	12.1	21.6	21.2	32.4	21.6
arthritis	-8.3	0.5	-0.6	2.7	1.6
health fair or poor	2.6	9.9	9.9	12.2	10.0
smoker	5.2	11.9	11.9	16.5	12.1
median	5.2	9.9	9.9	16.4	10.0

Table 6: Cost Saving (%) of Different Designs for $C_1/C_2 = 0.25$

Variable	One/HH	Opt. Integer	Half/HH	Opt. Fractional	Fractional ($n_g \geq 1$)
employment	1.3	9.5	9.5	9.5	9.5
unemployment	-14.6	3.1	2.2	5.7	3.1
income	0.3	5.9	5.9	10.4	6.1
language difficulties	7.0	18.9	18.1	25.2	18.9
arthritis	-14.6	0.3	-4.6	0.4	0.4
health fair or poor	-3.1	6.4	6.4	7.7	6.4
smoker	-0.3	8.5	8.5	8.5	8.5
median	-0.3	6.4	6.4	8.5	6.4

Table 7: Cost Saving (%) of Different Designs for $C_1/C_2 = 0.5$

Variable	One/HH	Opt. Integer	Half/HH	Opt. Fractional	Fractional ($n_g \geq 1$)
employment	-6.3	5.0	4.8	5.0	5.0
unemployment	-23.4	0.0	-2.9	0.5	0.5
income	-7.3	1.0	1.0	3.5	1.0
language difficulties	-0.1	15.2	13.8	15.2	15.2
arthritis	-23.3	0.0	-10.0	0.1	0.1
health fair or poor	-10.9	1.6	1.5	2.0	1.6
smoker	-8.0	3.9	3.7	3.9	3.9
median	-8.0	1.6	1.5	3.5	1.6